A Comparison of Item Parameter and Standard Error Recovery Across Different R Packages for Popular Unidimensional IRT Models

Taeyoung Kim and Insu Paek

Abstract With the advent of the free statistical language R, several item response theory (IRT) programs have been introduced as psychometric packages in R. These R programs have an advantage of a free open source over commercial software. However, in research and practical settings, the quality of results produced by free programs may be called into questions. The aim of this study is to provide information regarding the performance of those free R IRT software for the recovery item parameters and their standard errors. The study conducts a series of comparisons via simulations for popular unidimensional IRT models: the Rasch, 2-parameter logistic, 3-parameter logistic, generalized partial credit, and graded response models. The R IRT programs included in the present study are "eRm," "ltm," "mirt," "sirt," and "TAM." This study also reports convergence rates reported by both "eRm" and "ltm" and the elapsed times for the estimation of the models under different simulation conditions.

Keywords R IRT packages • eRm • ltm • mirt • TAM • sirt • Item parameter recovery

Many item response theory (IRT) estimation programs have been developed for the past years. Some commercial IRT programs are very widely used. For instance, PARSCALE (Muraki and Bock 1997) and MULTILOG (Thissen 1991) have frequently been used for research and in practice (Tao et al. 2014). Most notably, a free statistical language, R (R Core Team 2015), has provided several packages which have enabled researchers to conduct psychometric analyses. Rusch et al. (2013) outline the ongoing development of R packages in psychometrics, particularly in terms of breadth and depth in IRT.

T. Kim (✉)
State University of New York at Buffalo, Buffalo, NY 14228, USA
e-mail: tkim33@buffalo.edu

I. Paek
Florida State University, Tallahassee, FL 32306, USA
e-mail: ipaek@fsu.edu

© Springer International Publishing AG 2017

L.A. van der Ark et al. (eds.), *Quantitative Psychology*, Springer Proceedings in Mathematics & Statistics 196, DOI 10.1007/978-3-319-56294-0_36

As several IRT software have been introduced, comparisons among them for various model estimations have been studied. However, most studies have been limited to comparisons among commercial IRT packages. The earliest of these studies (e.g., Ree 1979) compared PARSCALE and MULTILOG under different population distributions for binary items. Later studies (e.g., DeMars 2002) encompassed a broad range of evaluations of these programs to polytomous items with Samejima's (1969) graded response model (GRM) and Masters' (1982) partial credit model (PCM).

Though previous studies have compared commercial and free IRT software (e.g., Pan and Zhang 2014), the IRT programs in R have not been rigorously evaluated in a systematic manner. Furthermore, most of the software evaluation studies have only investigated the recovery of item parameters in a variety of settings, and not of standard errors of item parameters. In this study, a comparison study was conducted via a series of simulations with popular unidimensional IRT models using five IRT programs in R, which are the Rasch model, the 2-parameter logistic (2-PL) and the 3-parameter logistic (3-PL) models, Muraki's (1992) generalized partial credit model (GPCM), and GRM, with respect to the recovery of item parameters and their standard errors.

The R IRT programs, at the time of the study, included the most updated versions[1] of "eRm" (extended Rasch modeling; Mair et al. 2015), "ltm" (latent trait models under IRT; Rizopoulos 2006), "mirt" (multidimensional item response theory; Chalmers 2012), "sirt" (supplementary item response theory models; Robitzsch 2015), and "TAM" (Test Analysis Modules; Kiefer et al. 2015). Except "ltm," the rest of the IRT programs in R were recently released.

1 Method

1.1 Conditions

We evaluated item parameter and standard error (SE) recovery under the following conditions for the dichotomous item response models: 2 (test forms) × 2 (sample sizes). Four conditions for the Rasch and 2-PL models were constructed by two test forms (test lengths of 25 and 50) and two different sample sizes (500 and 1000 examinees). For 3-PL model, the two test lengths were kept the same as those in the Rasch and 2-PL models, but sample sizes were increased to 2000 and 4000 based on preliminary analyses which have suggested a large sample size to avoid non-convergence issues. For the two polytomous models (GPCM and GRM), a single condition was considered: a large sample size of 5000 and a test length of six with each item having five categories. The purpose of using the large sample size was to avoid the zero frequency in some of the option(s), which presents challenges

[1] Note that this study used the latest version of each package available at the time of study: "eRm" (0.15–6; November 12, 2015), "ltm" (1.0–0; December 20, 2013), "TAM" (1.15–0; December 15, 2015), "sirt" (1.8–9; June 28, 2015), and "mirt" (1.15; January 21, 2016).

Table 1 Simulation design

Models	Test length (n)	Number of examinees (p)	Package(s) used
Rasch	25, 50	500, 1000	eRm, ltm
2-PL	25, 50	500, 1000	ltm, sirt, TAM, mirt
3-PL	25, 50	2000, 4000	ltm
GPCM/GRM	6	5000	ltm, mirt

in terms of evaluating the recovery of item parameters in reference to the true item parameters in the polytomous item response modeling. Also, the currently employed R IRT polytomous item response models do not provide a procedure to deal with this problem. While one package ("ltm") was evaluated for the 3-PL model, two packages ("eRm" and "ltm") and four packages ("ltm," "sirt," "TAM," and "mirt") were assessed for the Rasch model and the 2-PL model, respectively. For GPCM and GRM, "ltm" and "mirt" were evaluated. Table 1 encapsulates the simulation design in this study.

2 Data Generation

Item response data were generated following the standard IRT procedure. One thousand replications were made for each condition. Across all models, examinee ability (θ) was drawn from N(0, 1). True values of item parameters of dichotomous models were randomly drawn from logN(0, 0.5^2) for item discrimination or slope (a) parameters, N(0, 1) for item difficulty (b) parameters, and beta(5, 17) for the (pseudo) guessing (g) parameters. For the simulated tests, the true item difficulties ranged from 1.748 to 2.017 (mean = 0.088, SD = 1.024), the true discrimination ranged from 0.468 to 1.553 (mean = 1.000, SD = 1.72), while the true guessing ranged from 0.054 to 0.286 (mean = 0.185, SD = 0.056). For GRM, the same underlying distributions (i.e., logN(0, 0.5^2), N(0, 1)) were used again to generate true values of item discrimination parameters and step difficulty parameters, respectively. (It should be mentioned that the step difficulties (bs) were generated from N(0,1) and transformed into intercept parameters (d) by $d = ab$.) However, a simple item parameter set, which is not based on a random draw from the above distributions, was used for the GPCM data generation. This is because the current version of "mirt" does not use a popular GPCM parameterization, adopting a different parametrization from "ltm" with respect to the slope-intercept form in GPCM. (The current "mirt" GPCM parameterization is $a\theta - k$[2], where k is defined as a difference of adjacent intercept parameters, which is not conventionally used in the popular GPCM parameterization.) In this

[2] Note that "mirt" uses actually "+ intercept" but for consistency with the "ltm" expression, "–intercept" was used in this article.

regard, to make the metric transformation from "mirt" GPCM parametrization to the other usual slope-intercept form efficient, as for GPCM were either 1 or 2, and bs were −1, −0.5, 0.5, and 1 for the "mirt" GPCM calibration. Of note is that for the polytomous models, the recovery of the $a\theta - d$ parameterization was examined, while in the dichotomous models, the recovery of the $a(\theta - b)$ parameterization was investigated.

2.1 Recovery of Item Parameters and Their Standard Errors

The recovery of item parameters and their standard errors was examined after checking convergence of the model estimation. The evaluation criteria were absolute bias and root-mean-square errors (RMSEs). For the standard error recovery, the standard deviation of the parameter estimates was used as the (approximate) true value. With respect to standard error estimation, default methods provided by R IRT packages were used. "ltm" and "mirt" clearly delineated what the default standard error estimation method was. "ltm" reported standard errors using delta method under the usual IRT parameterization (i.e., $a(\theta - b)$ form). In "mirt" package, a variety of options for standard error computations, including "crossprod" which is the default, were available.

2.2 Convergence Check and Elapsed Time

This study reported estimation run times for all packages and convergence rates for "eRm" and "ltm" which provided a convergence indicator as part of the program run. Non-convergence rates and average elapsed estimation time per one data set are summarized in Table 2. Non-convergence rates shown in Table 2 represent the percentage of replication diagnosed by the program convergence indicator. Notably, the issue of convergence was critical in 3-PL model using "ltm." For the 3-PL model with "ltm," unreasonable estimates (e.g., very large unreasonable estimates) were sometimes observed despite the program reporting that there was no flag in the converge check.

3 Results

The results of this study, which excludes non-convergence replications, are summarized in Figs. 1, 2, 3 and 4: Rasch, 2-PL, 3-PL, and GRM, respectively. The summary measures (i.e., absolute bias, RMSE) in each of the figures represent averages across items. Our results suggest that absolute bias, and RMSE of item parameter estimates and their standard errors in "eRm," and "ltm" for the Rasch model, was nearly the same (see Fig. 1). We used a metric transformation to obtain equivalent parameter

Table 2 Average running time in minutes, average running time per iteration in seconds, and percentage of analyses that did not converge

Model	Sample size	Test length	Package	Time	Time/iter	Non-conv.
Rasch	500	25	eRm	17	1.02	0
			ltm	13	0.78	0
		50	eRm	41	2.46	0
			ltm	30	1.80	0
	1000	25	eRm	27	1.62	0
			ltm	20	1.20	0
		50	eRm	70	4.20	0
			ltm	56	3.36	0
2-PL	500	25	ltm	28	1.68	0
			sirt	16	0.96	NA
			TAM	27	1.62	NA
			mirt	16	0.96	NA
		50	ltm	56	3.36	0
			sirt	34	2.04	NA
			TAM	25	1.50	NA
			mirt	20	1.20	NA
	1000	25	ltm	36	2.16	0
			sirt	19	1.14	NA
			TAM	28	1.68	NA
			mirt	14	0.84	NA
		50	ltm	119	7.14	0
			sirt	41	2.46	NA
			TAM	44	2.64	NA
			mirt	33	1.98	NA
3-PL	2000	25	ltm	192	11.52	18.8
		50		365	21.90	19.3
	4000	25		480	28.80	22.8
		50		990	59.40	33.7

Note: Time = Average running time in minutes; Time/iter = Average running time per iteration in seconds; Non-conv. = Percentage of analyses that did not converge; and NA in Non-conv. represents convergence flag that was not available for those packages

estimates for the Rasch model, as “eRm” is based upon Rasch framework and uses sum-to-zero constraints for item difficulty estimates while this is not the case for “ltm” where a common item discrimination parameter is estimated.

Unlike the Rasch model, the 2-PL parameter recovery showed different performances across “ltm,” “TAM,” “sirt,” and “mirt.” Specifically, “TAM” showed relatively poor performance on point estimate recovery compared to other programs. For the SE recovery, “sirt” indicated poor performance as compared to the other programs. In general, “ltm” and “mirt” provided better results than the other two

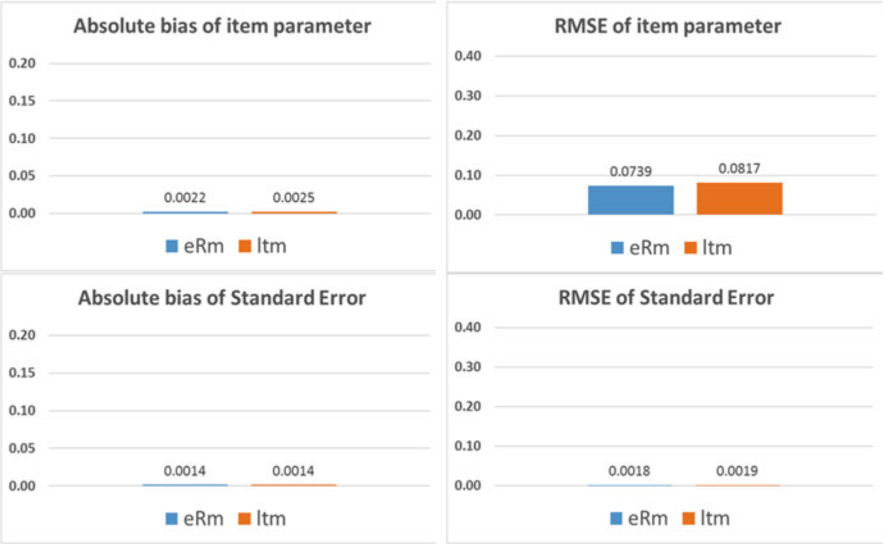

Fig. 1 Rasch result in case of n = 1000, p = 50

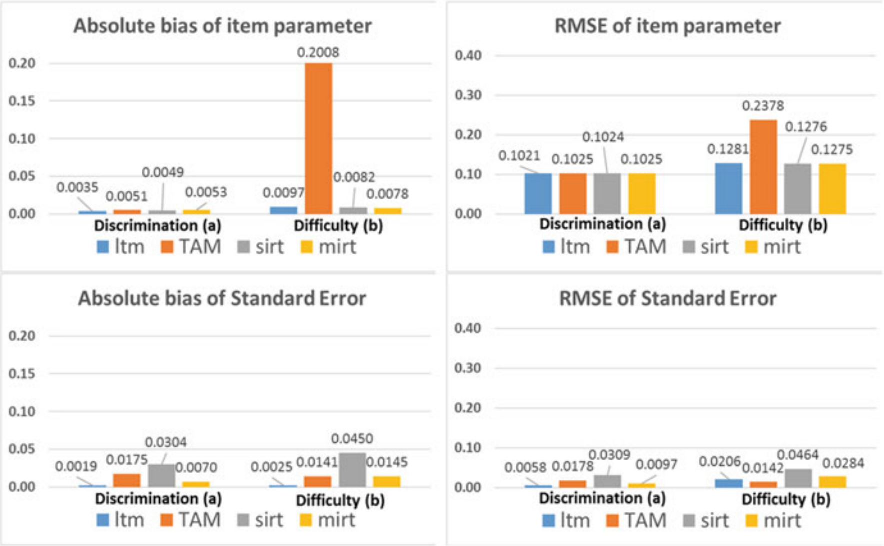

Fig. 2 2-PLM result in case of n=1000, p = 50

packages (see Fig. 2). For example, while the average RMSE of “ltm,” “TAM,” “sirt,” and “mirt” for discrimination parameter was 0.1021, 0.1025, 0.1024, and 0.1025, those for difficulty parameter were 0.1281, 0.2378, 0.1276, and 0.1275, respectively. “TAM” exhibited about twice average RMSE than the others. As well,

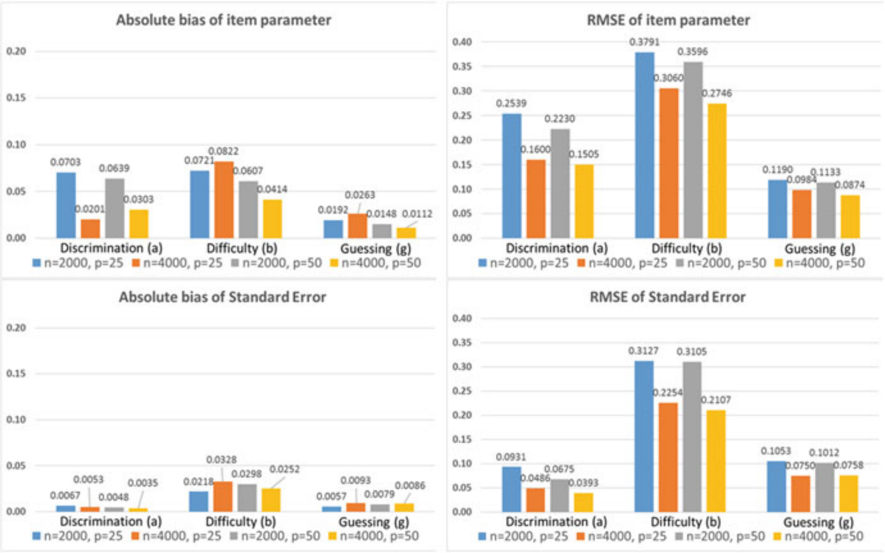

Fig. 3 3-PLM result

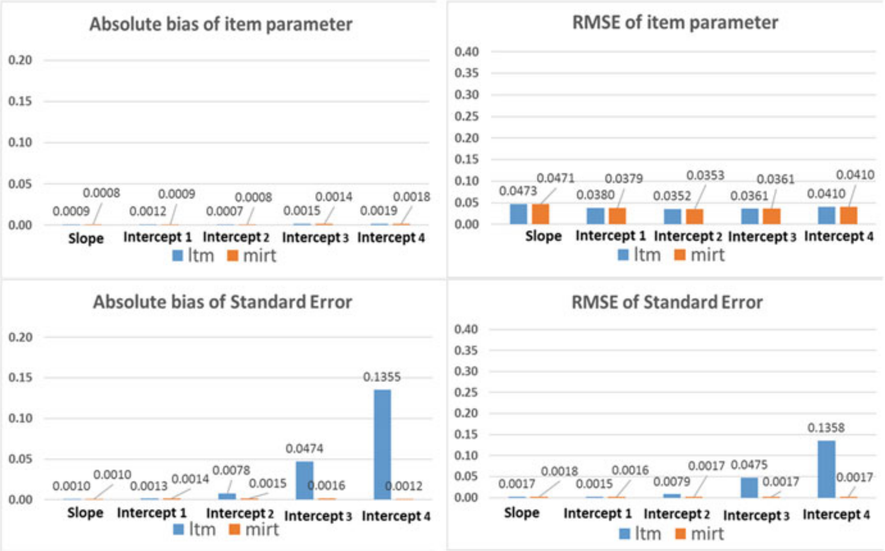

Fig. 4 GRM result

the average RMSE of "sirt" for the standard error of difficulty parameter was 0.0464, which was higher than those of "ltm," "TAM," and "mirt" (0.0206, 0.0142, and 0.0284, respectively).

In terms of the 3-PL model, only one package, "ltm," was used. As mentioned previously, the non-convergence rate was high in the estimation of the 3-PL model by “ltm,” which seems to be due to the lack of no item prior provision, especially for the low asymptote in the current “ltm” program. In addition, we observed that the convergence rate did not increase as the sample size increased (see Table 2). The RMSE values in the 3-PL model estimated by “ltm” were relatively high compared to the 2-PL model in general (see Fig. 3). In particular, while the average RMSE across four packages for discrimination parameter in the 2-PL model was 0.1024, that of the “ltm” 3-PL model was 0.1968 across different simulation conditions. This same pattern was also observed for difficulty parameter and standard error estimations of *a* and *b* parameters.

Both “ltm” and “mirt” provided either slope-intercept (i.e., $a\theta - d$ form) or conventional IRT parametrization (i.e., $a(\theta - b)$ form) for the polytomous item response models. However, as previously indicated, the intercept parameter in both programs for GPCM was not defined in the same manner. In “ltm,” the intercept is the usual intercept parameter itself (again, d in $a\theta - d$), while in “mirt,” it is defined sequentially (k in $a\theta - k$, which is the difference between adjacent intercepts). This different parameterization in both program made the comparison of SE challenging, although one may use a delta method. The current “mirt” program does not provide built-in standard error computation for $a\theta - d$ or $a(\theta - b)$. For this reason, only the evaluation of item parameter recovery was attempted in GPCM, and this study had more emphasis on GRM in terms of comparison of parameter and SE recovery for a polytomous model. The detailed results for GPCM are not presented here, but, overall, both “ltm” and “mirt” performed similarly in terms of the recovery of item parameters of GPCM. The RMSE values of all item parameters for both packages were very comparable (mean = 0.0425, *SD* = 0.0049 for “ltm”, and mean = 0.0421, *SD* = 0.0053 for “mirt”), while the absolute bias values were slightly smaller for “mirt” (mean = 0.0016, *SD* = 0.0005) than “ltm” (mean = 0.0036, SD = 0.0018), with respect to absolute bias. For the recovery of GRM, both “ltm” and “mirt” showed, again, comparable RMSE and absolute bias for the item parameter recovery, while the recovery of SEs noticeably differed across the two packages. In contrast to “ltm,” “mirt” exhibited stable performance with respect to the SE recovery. As illustrated in Fig. 4, while average RMSE of SE across item parameters (i.e., a slope and four intercept parameters) for “ltm” was 0.1945 (*SD* = 0.057), the corresponding quantity for “mirt” was 0.0017 (*SD* < 0.001). Finally, in terms of the program running time of the dichotomous response models (please see Table 2), “ltm” was faster than “eRm” in the Rasch model. For the 2-PL model, “mirt” was the fastest of the four packages. As expected, the elapsed time per replication for the 3-PL model by “ltm” was longest. With a sample size of 4000 and a test length of 50, it took nearly a minute for a single replication.

4 Discussion

This study evaluated the performance of free IRT programs in R regarding item parameter and its SE recovery. Because the programs are free, practitioners and researchers may consider those programs for classroom instruction, research, or other practical uses. In this regard, the results of this study provide a substantial amount of insight into the performance of five R IRT programs for popular unidimensional IRT models.

The ongoing continued development/update of some IRT programs in R and several limitations in this study warrant further research. The inclusion of currently popular commercial IRT software in the comparisons of these R IRT programs could provide even more insights, which would allow researchers and practitioners to recognize availability and potential utility of these R IRT packages. Of the current R IRT programs, the 3-PL model estimation by "mirt" requires further investigation. The high non-convergence rate and relatively weak performance of the 3-PL model estimation may be improved by employing item prior distributions, which are available in the "mirt" package. Finally, we suggest that users pay attention to the model parameterization used by each program, especially for the polytomous item response models. From the point of view of the consumer, R IRT program developers might consider providing more commonly used IRT parameterizations, as well as align the SEs of those parameters with common IRT parametrizations. This would prevent users being left to calculate SEs of those parameters manually (e.g., using the delta method by users).

References

P. Chalmers, mirt: a multidimensional item response theory package for the R environment. J. Stat. Softw. **48**(6), 1–29 (2012)

C. DeMars, Recovery of graded response and partial credit parameters in MULTILOG and PARSCALE. Paper presented at the annual meeting of American Educational Research Association, Chicago, IL, 2002

T. Kiefer, A. Robitzsch, M. Wu, TAM: Test Analysis Modules. R package version 1.15-0, 2015., http://CRAN.R-project.org/package=TAM

P. Mair, R. Hatzinger, M.J. Maier, eRm: Extended Rasch Modeling. R package version 0.15-6, 2015., http://CRAN.R-project.org/package=eRm

G.N. Masters, A Rasch model for partial credit scoring. Psychometrika **47**, 149–174 (1982)

E. Muraki, A generalized partial credit model: application of an EM algorithm. Appl. Psychol. Meas. **16**, 159–176 (1992)

E. Muraki, R.D. Bock, *PARSCALE 3: IRT Based Test Scoring and Item Analysis for Graded Items and Rating Scales [Computer Software]* (Scientific Software, Chicago, IL, 1997)

T. Pan, O. Zhang, A comparison of parameter recovery using different computer programs and the latent trait models R-Packages in estimating the graded response model. Paper Presented at the annual meeting of American Education Research Association, Philadelphia, PA, 2014

R Core Team, R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria, 2015. ISBN: 3-900051-07-0., http://www.R-project.org/

M.J. Ree, Estimating item characteristic curves. Appl. Psychol. Meas. **3**, 371–385 (1979)

D. Rizopoulos, ltm: an R package for latent variable modeling and item response theory analyses. J. Stat. Softw. **17**(5), 1–25 (2006)

A. Robitzsch, sirt: supplementary item response theory models. R package version 1.8-9, 2015., http://CRAN.R-project.org/package=sirt

T. Rusch, P. Mair, R. Hatzinger, Psychometrics with R: a review of CRAN packages for item response theory. Discussion Paper Series/Center for Empirical Research Methods, 2013/2. WU Vienna University of Economics and Business, Vienna, 2013

F. Samejima, Estimation of ability using a response pattern of graded scores. Psychometrika Monograph, No. 17, 1969

S. Tao, B. Sorenson, M. Simons, Y. Du, Item parameter recovery accuracy: Comparing PARSCALE, MULTILOG and flexMIRT. Paper presented at the 2014 National Council of Measurement in Education Annual Meeting, Philadelphia, PA, 2014

D. Thissen, *MULTILOG: Multiple Category Item Analysis and Test Scoring Using Item Response Theory [Computer Software]* (Scientific Software, Chicago, IL, 1991)